下载APP

18 | 马尔可夫链：你的未来，只取决于你当下做什么

2021-09-08 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 14:08 大小 12.95M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

上节课讲了拉斯维加斯和蒙特卡洛算法，结合前面的基础算法你会发现这些算法的特点是来解决某个时间点的问题，但没有解决那些和时间先后次序相关的预测问题。

我们现实生活当中其实充满了很多和事情顺序相关的过程。也就是说一件事情发生后会影响另外一件事情的结果，而这些事往往是按照某一个规律次序发生的。今天我们就来聊聊和时间序列预测相关的一个算法：马尔可夫链。

马尔可夫链专门研究在现实生活当中这一系列的事件，**找到它们的内部运行规律，从而预测当这一系列事件达到平衡的时候，当前状态的下一步最可能发生的情况。**这样我们就可以知道，当一件事情发生的时候，未来有多大可能会发生另一件事情。

马尔可夫链算法定义与场景

马尔可夫链因俄国数学家安德烈·马尔可夫得名，它的定义是：状态空间中经过从一个状态到另一个状态的转换的随机过程。该过程要求具备“无记忆”的性质，也就说下一状态的概率分布只能由当前状态决定，在时间序列中它前面的事件均与之无关。

看完这个定义，你会不会有种一头雾水的感觉？我用一个简单的例子给你解释一下。天气就是一个状态，比如昨天是阴天，今天是晴天。如果今天的天气只和昨天天气有关（也就是和昨天之前的任何一天的天气都没有关系），那么，天气的这个系统就是一个符合马尔可夫链的完备系统，我们就可以通过今天的天气来预测明天的天气，甚至预测未来 1 个月、1 年的天气。

而马尔可夫同学在最开始建立马尔可夫链的时候，最著名的应用就是从一份俄罗斯诗歌作品中数出几千个两字符对，他使用这些两字符对计算了每个字符出现的概率。也就是在这诗歌里我们给一个字符，就能预测关于下一个字符可能是什么。

通过这个方法，马尔可夫可以模拟这个诗歌一个任意的长字符序列，这就是马尔可夫链。所以，如果这个概率很准确的话，在这个诗歌里看到前面的这句话，你大概率就知道下一句会用什么样的词和句子了。我们甚至可以让计算机自己创造作品，例如现就就有计算机算法自己学习巴赫的作品，然后生成一首巴赫的曲子，感觉神韵还很像，你可以点击这里去这里听一下。

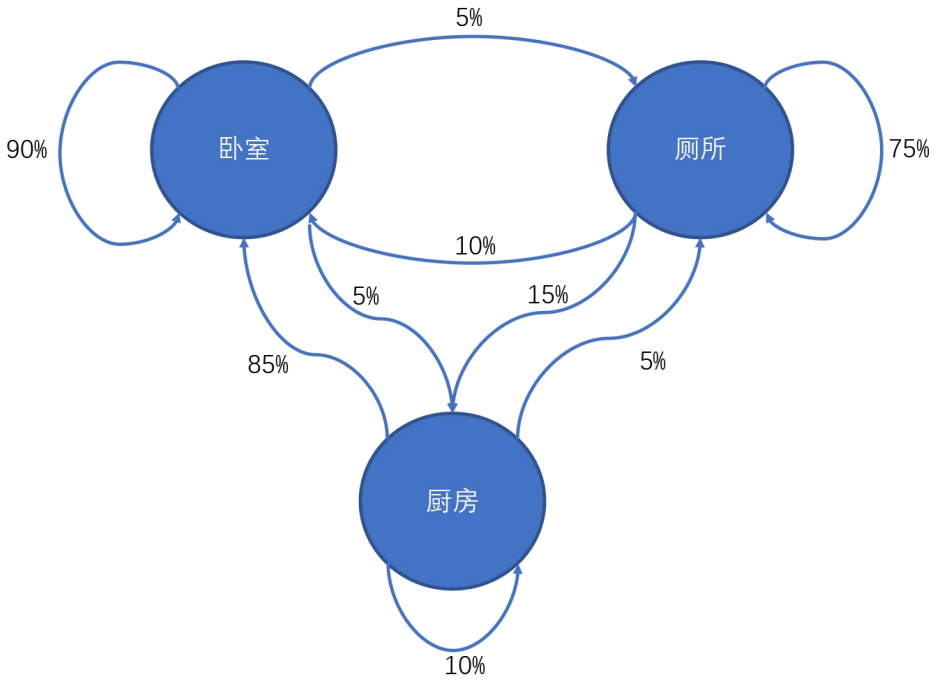
有非常多著名的算法例子都使用了马尔可夫链。比如著名的谷歌创始人拉里·佩奇和谢尔盖·布林在 1998 年提出的谷歌搜索最核心的网页排序算法 PageRank 就是由马尔可夫链定义的，而这个算法造就了整个谷歌在搜索引擎里的霸主地位，至今为止你去谷歌搜索内容的准确率还是远高于其他搜索引擎的。

又比如在经济金融领域，詹姆斯.D. 汉密尔顿在 1989 年用马尔可夫链来对高 GDP 增长速度时期与低 GDP 增长速度时期（也就是经济扩张与紧缩）的转换进行建模，帮助美国在经济萧条中对 GDP 的恢复情况进行预测，直到今天马尔科夫链依然是经济学当中预测一个国家 GDP 重要方法之一。

马尔可夫链算法举例

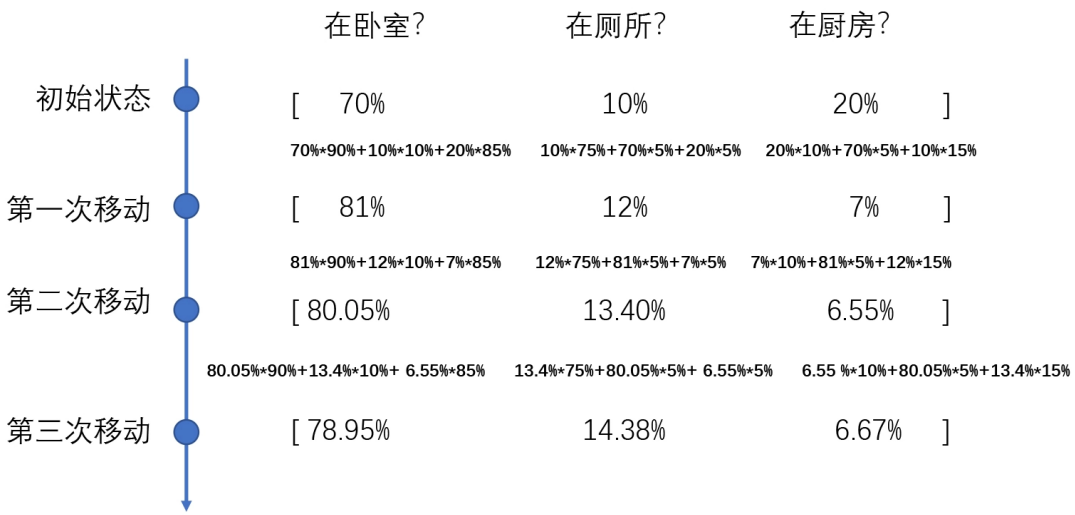
马尔可夫链这么厉害，我们再更具体地来看一下，马尔可夫链的原理是什么，以及我们可以怎么用。

假设你现在住在一居室里，这房子里有卫生间、卧室、厨房，你把你从一个房间到另一个房间的概率统计了一下。例如你现在在卫生间，你有 75% 的概率会留在卫生间，有 10% 的概率会走到卧室，有 5% 的概率会走到厨房。你把你从每一个房间到另一个房间的概率都统计清楚了，这就可以形成一个马尔可夫链，如下图所示。

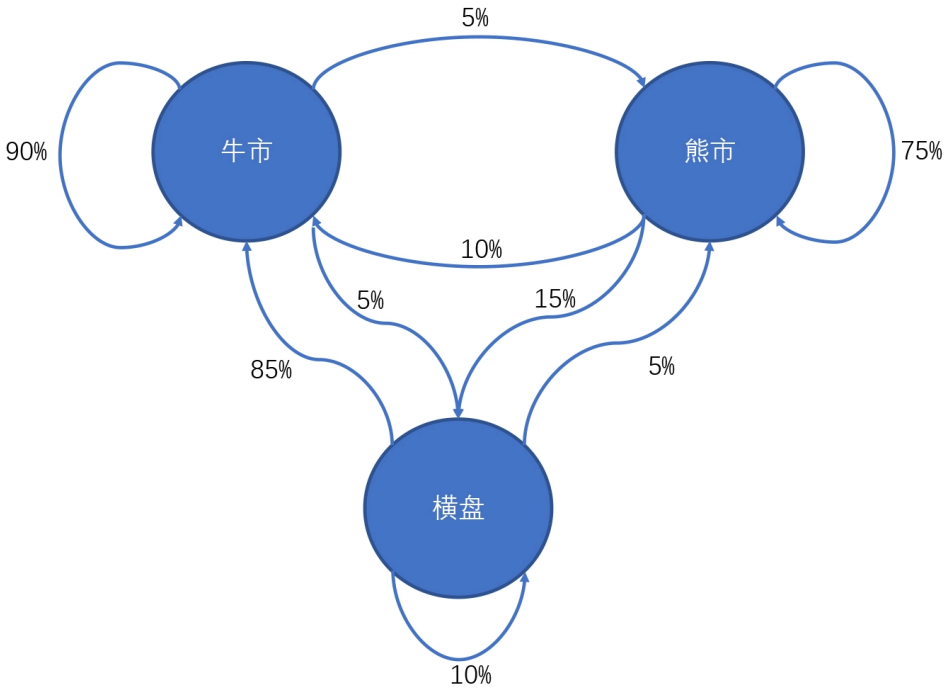


上图中的箭头和旁边的数字表示从一个状态转移到另一个状态的概率，例如，从卧室到卧室是 90%，从卧室到厨房概率 5%，从卧室到厕所概率 5%。我们可以通过大量的统计分析或者算法预测来完善这个马尔可夫链里边的概率。最终我们会看到你要去哪一个房间，和你最开始在哪个房间其实并没有关系，只和你上一个所处的房间有关系的时候，这时就可以通过马尔可夫链来计算你的个人行为的长期趋势了。

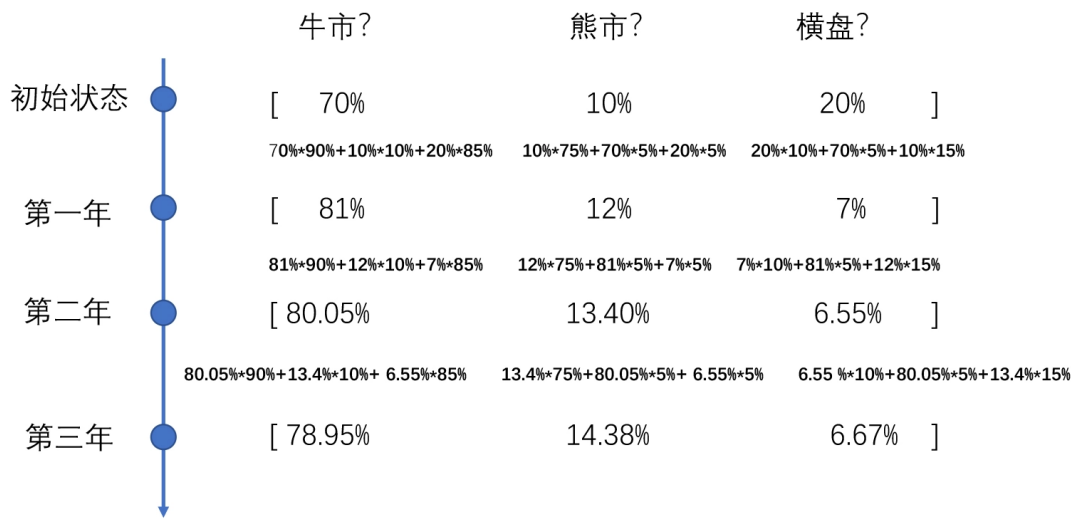
比如假设我初始的时候，在卧室、厕所、厨房的概率分别是 70%、10%、20%，那么我可以预测移动三次之后，我去每个房间的概率是多少，就像下面这个图展示出来的。这个图就叫做“转移矩阵”，有了它我们就可以对整个我在房间里的行为有一个规律性的判断，例如，从前面的规律和初始概率，从下面的推算会发现，我在厕所的时间越来越长……



这个时候，你可能会问了：“预测房间这件事情没有什么实际意义啊？”那什么才会更有实际意义呢？我们可以把上面的三个房间替换成股市的三个状态，分别是牛市，熊市和横盘，就变成现在下图的样子。



再计算一下股市的“转移矩阵”，如下图。



这个时候我们是不是就可以根据现在的股市的状态，看将来会是熊市还是牛市了呢？其实在金融行业里，这就是马尔可夫链最常用的一个案例。只不过在股市里，每个状态转换的概率要复杂得多，计算机很难计算出来。于是，结合前面我们讲的蒙特卡洛算法，就有了大名鼎鼎的**马尔可夫链蒙特卡罗算法（MCMC）**。

MCMC 由梅特罗波利斯于 1953 年在洛斯阿拉莫斯国家实验室提出，本质上是将马尔可夫链用于对蒙特卡洛方法的计算过程中。那时美国洛斯阿拉莫斯是当时少数几个拥有大规模计算机的城市，梅特罗波利斯利用这种计算优势，在蒙特卡洛方法的基础上引入马尔可夫链，用于模拟某种液体在气化之后的平衡状态。

1984 年 Stuart 和 Donald Geman 兄弟对吉布斯采样进行了描述，形成了我们今天所熟悉的版本，这些算法在自动化交易和临床医学上都有很多的应用，比如当测试数量趋向无穷时，MCMC 方法可以将病人症状与方剂药效持续配对，直到最后完全逼近出一个虚拟的人体模型作为状态观测器，并总结按照输入输出关系模型反馈给药的规律。

同样，在互联网公司里，要去做一些推荐的时候，也会用到马尔可夫链的一些算法。我们去浏览网站，其实无外乎就是在浏览、购买、收藏商品。其实这些行为也可以变成和上述移动房间类似的马尔可夫链形式，这样我们可以根据每一个不同的行为状态来预测下一步用户可能会做什么，然后这个时候我们给用户最方便的一些行为指导和预测，就可以促进用户的购买。

这也就是你在淘宝上看到的“猜你喜欢”和首页推荐列表里其中的一个核心算法。我们可以根据用户的“特征喜好状态转移矩阵”，得出用户可能在下个时刻的操作列表，然后把它做成推荐列表。最后将多个推荐列表进行其他算法的加权融合，得出最终的列表结果。

算法应用场景及展望

马尔可夫链的应用非常广泛，例如天气的预测、食品销售的预测、GDP 的涨幅预测、企业人员的变动预测等等问题，都可以通过马尔可夫链来解决。当这些复杂系统某些条件进行变化的时候，马尔可夫链就可以根据前面的转移矩阵先推算未来最可能的状态，从而对政府和企业的决策产生非常重大的影响。

在人工智能领域里，Siri 里面自然语言的识别就经常会用到马尔可夫链的预测。因为我们说的话里上一句和下一句，上一个词和下一个词，基本上也是遵循马尔可夫链的规律的。所以**我们会通过马尔可夫链来修正计算机识别的一些问题。**

比如说使用马尔可夫的算法根据前一个状态识别出下一个状态正确单词，就会比单独识别某一个单词准确率高得多。这样根据上下文你讨论着某一个职位“Vocation”，最后 Siri 就不会理解成你要去度假旅游“Vacation”，最终给你的回答和答案也更加准确了。

又比如说，我们在自动驾驶识别前面路况时，到底识别出是道路、天空还是学校？通常做法我们会使用它的颜色用项数据标记，但这个时候可能就会有各种各样复杂的变量和因素，如果只看局部不加上下文马尔可夫链算法的判断，就会决策出现一些问题，例如在天空里面识别出一段公路或者是把前面的卡车当成了天空。

其实我们知道在图像当中，前后时间比较近的这些像素通常都使用同一个物体，这个时候我们就可以把在前后时间相近、特征相邻的像素识别为相同的东西，就不容易识别错误，避免突然在空荡的马路上停车，或者看着一个卡车当作天空直接撞上去了。

现在有人把马尔可夫链用于分析生物的 DNA 序列，还有人用马尔可夫链算法来预测双色球和其他彩票，以及做炒币的自动化交易，以及我们在前面讲的用马尔可夫链来作曲等等。

总之在跟序列相关的反馈机制预测问题上，马尔可夫链是非常有帮助的。不过**马尔可夫链和其他数据算法联系非常紧密，它的预测结果好坏其实都依赖于我们刚才提到的概率转移矩阵是否准确，而这个概率转移矩阵的准确性最后又依赖于算法估算方法的合理性。**所以

马尔可夫链要算准确，需要建立在前面我们提到的基本算法（回归、分类、聚类、关联等）预测概率的准确性上。

马尔可夫链除了受到其他算法的限制之外，本身也有它的局限性。马尔可夫链的假设是后一个状态值和前一个状态相关，而和更靠前的状态无关。而这个假设在一些情况下是不太符合实际的。例如我买一个品牌的衣服，发现质量特别差，可能未来买 100 次衣服我都不会去考虑这个衣服品牌了。所以**虽然马尔可夫链应用这么广泛，是不是要用马尔可夫链算法，还是要结合具体业务场景。**

小结

这节课的内容到这里也就接近尾声了，最后我来给你总结一下。今天我们主要讲了马尔可夫链，它是非常著名的有序状态相关算法，可以帮助我们从一系列的事件里面找到内部运行规律，从而预测未来的情况。从 Google 搜索引擎到预测股市，从语音识别到自动驾驶，甚至是自动作曲作画、预测国家 GDP 的增长都可以使用马尔可夫链算法。

其实，在我们工作和日常生活当中也有很多“马尔可夫链”：你现在的状态其实大部分都是由你上一个状态决定，没有人会走背字一直失败，也没有人能幸运到一直成功。

你可以仔细想想，真正的失败，很多时候都是自己遇到失败后从此一蹶不振，走不出来失败的这个状态才造成的。“没有迈不过去的坎”这句话用马尔可夫链的视角来看，那就是现在自己的状态，只和自己上一个状态相关，和整体无关。所以吸取完教训后，调整好现在的心态，用现在去影响你的未来。

我特别喜欢《飘》电影结尾郝思嘉说的那句话，我觉得它诠释了“马尔可夫链”在生活哲学中的真谛：“Tomorrow is another day”——**你的未来只取决于你当下在做什么，而不是过去你曾经做过什么，毕竟“明天，是新的一天”。**

数据给你一双看透本质的眼睛。让我们一起从数据世界里参悟到一些人生哲理，让数据“驱动”我们的生活。

课后思考

你曾经遇到过哪些事情是只由上一个状态（而不是过去状态）来决定你下一个状态的么？最后你收获到你想要的了吗？它能总结成一个马尔可夫链么？分享出来我们一块过过招？

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 2 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 17 | 蒙特卡洛与拉斯维加斯：有限时间内如何获得最优解？

下一篇 19 | 协同过滤：你看到的短视频都是集体智慧的结晶

精选留言 (3)

写留言

nicole_yaoyao

2021-09-15

人生就是一环接一环的马尔科夫链，强者愈强，大数定律，但是也总会有小数陷阱，所以强者也会遭遇滑铁卢，关键是，如果你能振作起来，整体就是向上的；弱者也不一定越弱，如果在恰当的时机抓住了机会，也能够逆天改命

展开

1

艺霖子

2021-09-10

明明是数据分析课，偏偏带上了我最爱的心理学，数据分析课治愈了我---

进化菌

2021-09-08

马尔可夫链算法，有点像游戏里的打怪升级，当前的努力打怪，才有可能有机会挑战更高级别的boss。

在我们的生活里，我想到的是种花，你得先播种，然后经常性松土施肥浇水，花儿才会一天天长大，也才有可能开出花儿来。...

展开